Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 4 (2015), pp. 1899-1906

http://www.ripublication.com

A General Critical Review on Privacy Preserving Data Mining Techniques

G.Manikandan[1]

Assistant Professor, School of Computing, SASTRA University, Thanjavur, India
manikandan@it.sastra.edu

N.Sairam[2]

Associate Dean / IT, School of Computing, SASTRA University, Thanjavur, India
sairam@cse.sastra.edu

M.Sathya Priya[3], Sri Radha Madhuri[4]

Student / B.Tech ICT, School of Computing, SASTRA University, Thanjavur, India
[3]m.sathya1994@gmail.com, [4]radhashreyas@gmail.com

Abstract

With the widespread quantity of records stored in various storage repositories it is exceptionally significant to craft a dominant and valuable mechanism for analyzing such data for retrieving the appealing patterns that could help in decision making. Data mining is a predominant technique which is used to extract the valuable details and patterns from the hefty repositories. People today turn out to be well conscious of the privacy seepage of their perceptive data and are extremely averse to share their information. The key area of apprehension is that non-sensitive data could convey insightful information such as facts and interesting patterns. Privacy preserving data mining (PPDM) is a new track in data mining which focus on hiding an individual's characteristics without compromising the data usability. The primary idea of privacy preserving data mining was to perform data mining on confidential data. A number of schemes and routines have been utilized for this purpose. Data sanitization and extracting the data mining result from the modified data were the key issues to be addressed. The aim of this paper is to provide a

complete review on approaches and techniques based on different aspects for ensuring privacy in data mining and points out their advantages and disadvantages.

Keywords: Data Privacy; Data Perturbation; Data Accuracy; Data Utility; Data sanitization.

1. INTRODUCTION

In these days it is easier for organizations to store any small piece of information obtained from the current behavior of their clients as hardware costs goes down day by day. With the phenomenal growth of internet and relevant technologies in the precedent years has resulted in a wealth of information that can be used by number of commercial companies and various government agencies. Data holders regularly look for the better use of data they possess and make use of data mining tools to pull out constructive facts and patterns from the data. Due to the extensive on hand data and looming need, data are curved into handy information. The knowledge acquired could be utilized for a mixture of applications ranging from customer retention, market analysis and credit card fraud detection.

Data mining is the procedure of ascertaining interesting knowledge from heterogeneous repositories. It is an integration of various fields that includes data visualization, information retrieval, high performance computing, database technology and machine learning. Information can be extracted in different dimensions using data mining. The revealed facts can be used in appropriate decision making and efficient query processing [1].

Data mining is considered as one of the most capable interdisciplinary development in the information industry. Data mining, with its guarantee to proficiently discovering priceless information from outsized databases, is predominantly susceptible to misuse. So, there may be a disagreement among data mining and privacy [2].

In spite of its usefulness, many data providers are unenthusiastic to afford their data for data mining for the trepidation of violating individual privacy. Data mining has been viewed as a peril to privacy of data that is owned by the industry. It integrates privacy as a purposeful component for the gained information and allows an organization to use the stored data to develop new relations to get better efficiency. Hence Privacy preserving data mining turn out to be a significant field to explore.

Privacy Preserving Data Mining (PPDM) is a new research direction in data mining. Its main objective is to acquire the exact results of data mining while protecting the sensitive information from seeping out of the mining process from the sanitized data. Numerous techniques have been developed for privacy preserving data mining [3]. Many privacy preserving techniques are using some form of data alteration to accomplish privacy. These techniques focus on data distortion, data reconstruction and data encryption to achieve privacy. The implementation of PPDM techniques has turn out to be a great demand at present. High data eminence with isolation is the foremost prerequisite of good quality privacy preserving techniques.

The objective of this article is to present the analysis on privacy preserving procedure which is exceedingly supportive in mining large data sets with realistic competence and security. The paper is organized as follows. In Section 2, we give the privacy issues in data mining. In Section 3, we describe Privacy Preserving techniques with their merits and demerits. A tabular comparison of different techniques is shown in section 4 and finally we conclude in Section 5.

2. PRIVACY ISSUES RELATED TO DATA MINING

Organizations make use of a range of Data mining Algorithms to mine appealing patterns that support their daily schedule. These data may encompass interested information with reference to individuals. A malevolent data excavator can gain knowledge of perceptive data of a certain person by means of re-identification from an uncovered data source. The amalgamation of supplementary attributes also assists the interloper to recognize the insightful data values [4 -5].

In recent epoch, several studies have been made to guarantee that the insightful information of persons cannot be easily recognized. Privacy is defined as the quality of being isolated from the vision of others. Privacy is a subject of apprehension because it may have undesirable effects on individual's life. Privacy is not desecrated till a person feels that his private information is being used unconstructively. Once it is exposed then it cannot be prevented from being misused [6].

These models assume that the data contains two different categories of attributes namely Sensitive attributes and Quasi Identifier attributes. Attributes like Date of birth, Zip code, Sex are examples for quasi identifier attributes where as attributes like Type of Disease, Patient name are examples for sensitive attributes. It is also understood that each record in a table correspond to an individual entity and no two records refer to the same entity.

We consider a situation in which two or more organizations are in possession of their private files desires to execute a data mining algorithm on the unification of their files without enlightening any concealed facts. For instance, consider two different medical institutions that desire to carry out a joint study without violating the privacy of their patients. In this situation it is mandatory to shield the insightful information ensuring its use for upcoming research. Organizations recognize that the mix of their data offer mutual advantage, but no one is prepared to divulge its database. For this reason a variety of privacy preserving techniques are used along with data mining algorithm to shield the mining of responsive information all through the data extraction.

3. PRIVACY PRESERVING TECHNIQUES

In the last few years, numerous approaches have been projected in the area of privacy preserving data mining. These techniques can be classified as Cryptographic methods, Query auditing and Data modification methods based on the diverse protection methods as shown in Fig-1. In data modification techniques original data is altered

before issuing it to the users. Data is customized in such a way that the privacy is preserved in the published data set. Several data modification methods are suggested including noise addition[7][8][9], data swapping[10][11], aggregation[12][13], suppression [14][15] and signal transformation[16][17].

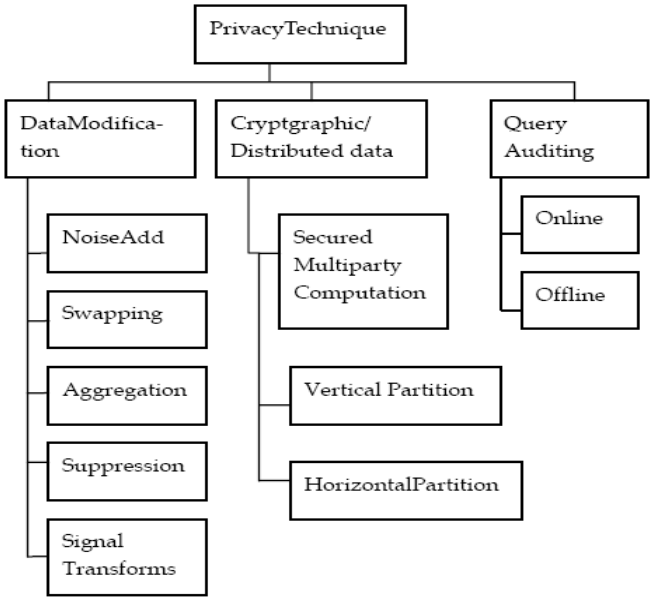

Fig-1: Classification of Privacy Preserving Data Mining Techniques.

3.1 *k*-Anonymization Techniques

Samarati and Sweeney introduced the concept of *k*-anonymization in [18-19]. A database is said to be k-anonymous if there exists at least *k records* in the database having the same values for the quasi-identifier attributes. In order to guard the sensitive information in the database *D*, before relinquishing D to the public, D is transformed into a fresh dataset D' that ensures the *k*-anonymity for a sensitive attribute by doing some value generalizations on quasi-identifier attributes. Therefore, the level of uncertainty of the sensible elements is at least 1/*k*.

3.2 Query Auditing:

Query auditing process preserve privacy by altering or changing the answers of a query. [20-22]. Query auditing can be classified into two categories namely offline auditing and online auditing. In offline auditing the auditor's responsibility is to unravel a system of linear equations to verify whether the reply to these queries can be applied to inimitably deduce some private value. The role of an online auditor is to check whether the new query should be replied or deprived of in order to avert privacy breach. The utility of the data set is reduced by these methods if there are too many denials. On the contrary lesser denials increase the utility of the data set but privacy is sacrificed.

3.3 Cryptographic technique:
The cryptographic approaches usually promise elevated data privacy. Many Cryptography-based methods have been projected in the framework of privacy preserving data mining algorithms. These methods [23] encrypt the data set by means of a set of algorithms such as secured multiparty computation (SMC). SMC techniques divulge only the ultimate result of the computation. Usually SMC techniques are implemented in distributed scenarios. For centralized database SMC protocol is implemented by partitioning the database either vertically or horizontally [24-26]. Cryptographic methods are less effective for bigger data sets where data usage is of a greater concern.

3.4 Fuzzy Logic:
Fuzzy set has dominated the area of knowledge representation and translation for a long period of time. Theory of fuzzy sets has been widely accepted in various application domains like Data mining, Pattern Recognition and machine intelligence. Fuzzy set theory assist in representing uncertainty and approximation. Human reasoning requires complete information for making appropriate decisions and it fails in the case of ambiguous or partial information's. Fuzzification is used to transform the data points in the crisp set to fuzzy set [27-28]. Each fuzzy set is assigned with a unique linguistic variable which correspond to a concept that is computable either subjectively or objectively. Fuzzy sets can be used for preserving privacy in Data mining [29-30]. Fuzzy logic relies on a membership function to sanitize the original data. Data Utility mainly depends on the type of membership function used.

3.5 Data Transformation: -
Oliveira and Zaiane revised a class of geometric data transformation methods that alter numerical values by applying scaling, shearing, translation functions or by the combining all the above transformations. In translation a constant additive noise is added to the original data to generate the modified data. A constant multiplicative noise is added for data sanitization in the case of shearing and scaling. Geometric data transformation method was intended to promise valid data mining results and to guarantee privacy requirements of data owners as well. The advantage of this approach is that it allows end users to make use of their own techniques to be utilized ahead of the mining procedure to gratify the privacy constraint on the data [31-33].

4. COMPARATIVE STUDY

The data set used in this work is the Pima Indians diabetic data set which belongs to National Institute of Diabetes and Digestive and Kidney Diseases available on UCI Machine Learning Repository [34]. It has 768 records where each record contains information about at least 21 year old female person of Pima Indian heritage. Age attribute is used for clustering purpose. The above proposed approaches have been implemented using JAVA Programming language and the resulting observations are tested in Intel core i5 processor with 4GB RAM and Windows 8 operating system. We have tabulated few of the results below for comparison. From our experimental

results it is evident that the original data cannot be inferred from the modified data which is shown in Table 1.

Table 1 Original and Modified Data for a Synthetic Data Set containing 12 sample values.

Original Data	Modified Data after applying PPDM Techniques				
	k-anonymity one level	Fuzzy logic	Translation Noise=10	Scaling Noise=10	Shearing Noise=10
02	0*	0	12	20	22
04	0*	0.0102	14	40	44
10	1*	0.1632	20	100	110
12	1*	0.2551	22	120	132
03	0*	0.0025	13	30	33
20	2*	0.0749	30	200	220
30	3*	1	40	300	330
11	1*	0.2066	21	110	121
25	2*	0.9362	35	250	275
12	1*	0.2551	22	120	132

In Table 2 we summarize the comparative work of the privacy preservation techniques on the basis of evaluation criteria which we mentioned above.

Table 2: Privacy Preserving Techniques – Comparative Study

S.No	PPDM Approach	Privacy Level	Data Type	Data Utility
1	k-Anonymity	Yes	Numeric Categorical	Less
2	Cryptographic	Yes	Numeric Categorical	Less
3	Query Auditing	Yes	Numeric Categorical	Depends on Query Denials
4	Translation	Yes	Numeric	Depends on the added noise
5	Rotation	Yes	Numeric	Depends on the added noise
6	Shearing	Yes	Numeric	Depends on the added noise
7	Fuzzy logic	Yes	Numeric	Depends on the added noise

5. **CONCLUSION AND RESEARCH DIRECTIONS**

We have reviewed various techniques used for achieving privacy in data mining. The effectiveness of these techniques is evaluated using a set of criteria like privacy level, data type and data utility. As of now none of the existing PPDM algorithms provides a complete solution for achieving better privacy and data utility. There are abundant research directions all along the way to quantify upcoming PPDM algorithms. One among them is to develop a unified model for evaluating and comparing various PPDM algorithms by using different metrics. It is also important to design and develop standard datasets for examining the entire set of PPDM algorithms.

References:

[1] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kauffman Publishers, 2006.

[2] G.K.Gupta, *Introduction to Data Mining with Case Studies*, Prentice Hall of India, 2008.

[3] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu; Philip S. Yu, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, Chapman and Hall, 2010

[4] Margaret.H.Dunham, Data Mining: *Introductory and advanced topics*, Pearson Education, 2003.

[5] K.P.Soman, Shyam Diwakar, V.Ajay, *Insight into Data Mining: Theory and Practice*, Prentice Hall of India, 2006

[6] Jaideep Vaidya Christopher W. Clifton, Yu Michael Zhu, *Privacy Preserving Dat Mining*, Springer, First Edition, 2005.

[7] Agrawal R., Srikant R., ``*Privacy Preserving Data Mining.*, '' In the Proceedings of the ACM SIGMOD Conference. 2000.

[8] K.Muralidhar., R.Sarathy., ``*A General additive data perturbation method for data base security.*, '' Journal of Management Science., 45(10):1399-1415, 2002.

[9] Agrawal D. Aggarwal C.C. " *On the Design and Quantification of Privacy Preserving Data mining algorithms*." ACM PODS Conference, 2002.

[10] Fienberg S.E. and McIntyre J. "*Data Swapping:Variations on a theme by Dalenius and Reiss.*" In Journal of Official Statistics, 21:309-323, 2005.

[11] Muralidhar K. and Sarathy R. " *Data Shuffling- a new masking approach for numerical data*." Management Science, forthcoming, 2006.

[12] Y.Li, S.Zhu, L.Wang, and S.Jajodia " *A privacy-enhanced microaggregation method*" In Poc. Of 2nd International Symposium on Foundations of Information and Knowledge Systems, pp148-159, 2002

[13] V.S. Iyengar."*Transforming data to satisfy privacy constraints*" In Proc. of SIGKDD'02, Edmonton, Alberta, Canada, 2002.

[14] A.A.Hintoglu and Y.Saygin. " *Suppressing microdata to prevent probabilistic classification based inference*" In Proc. of Secure Data Management, 2nd VLDB workshop, SDM 2005 pp155- 169, Trondheim Norway 2005.

[15] S.Rizvi, J.R. Harista "*Maintaining data privacy in association rule mining*" In Proc. of 28th VLDB Conference, pp682-693, Honk Kong China, 2002.

[16] Shuting Xu., Shuhua Lai, "*Fast Fourier Transform based data perturbation method for privacy protection*" In Proc. of IEEE conference on Intelligence and Security Informatics, New Brunswick New Jersey, May 2007.

[17] Shibanth Mukharjee., Zhiyuan Chen., Arya Gangopadhyay "*A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms*" The VLDB journal 2006

[18] Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE) **13**(6), 1010–1027 (2001). DOI HTTP://doi.ieeecomputersociety.org/10.1109/69.971193

[19] Sweeney, L.:"*Achieving k-anonymity privacy protection using generalization and suppression"*, International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), 571– 588 (2002)

[20] Blum A., Dwok C., McSherry F., Nissim K. " *Practical Privacy The SuLQ Framework*" ACM PODS Conference, 2005

[21] Kenthapadi K., Mishra N., Nissim K., "*Simulatable Auditing*"ACM PODS Conference 2005.

[22] Nabar S. Marthi B., Kenthapadi K., Mishra N., Motwani R., "*Towards Robustness in Querry Auditing"* VLDB Conference, 2006.

[23] Pinkas B."*Cryptographic Techniques for Privacy-Preserving Data Mining*" ACM SIGKDD Explorations, 4(2), 2002

[24] Lindell Y., Pinkas B."*Privacy preserving Data Mining*"CRYPTO 2000.

[25] Yu H., Jiang X., Vaidya J."*Privacy Preserving SVM using nonlinear Kernels on Horizontally partitioned Data*. SAC Conference, 2006.

[26] Yu.H., Vaidya J., Jiang X."*Privacy preserving SVM Classification on vertically partitioned data"* PAKDD conference, 2006.

[27] Zadeh L "*Fuzzy sets*", Inf. Control. Vol.8, PP, 338 – 353, 1965.

[28] Timothy J. Ross "*Fuzzy Logic with Engineering Applications*", McGraw Hill International Editions, 1997.

[29] V.Vallikumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana, BVS Avadhani "*Fuzzy based approach for privacy preserving publication of data*", IJCSNS, Vol.8 No.1, January 2008.

[30] Karthikeyan B, Manikandan G et al. "*A fuzzy based approach for privacy preserving clustering"*, Journal of Theoretical and applied information Technology, vol 32(2), 118–122, 2011.

[31]. Rajalaxmi R R, and Natarajan A M, "*An effective data transformation approach for privacy preserving clustering"*, Journal of Computer Science, vol 4(4), 320–326, 2008.

[32]. Manikandan G, Sairam N et al., "*Privacy preserving clustering by shearing based data transformation"*, Proceedings of International Conference on Computing and Control Engineering, 2012.

[33] Stanley R. M. Oliveira, and Osmar R. Zaiane, *"Revisiting Privacy Preserving Clustering by Data Transformation, "* Journal of Information and Data Management, vol. 1, no. 1, 2010.

[34] UCI Data Repository http://archive.ics.uci.edu/ml/datasets.html